Sampling Methods

David A. Hughes, Ph.D.

Auburn University at Montgomery

david.hughes@aum.edu

March 30, 2022

Why sampling?

- We want to draw inferences about populations of interest.
- We don't have infinite resources.
- Under the right circumstances, it's pretty accurate.

Sampling basics

- What's the target population?
- How could we construct a sampling frame for it?
- How can we select samples/elements from the sampling frame?
- How accurate or generalizable are these samples?

An example: Weight of dogs

- Suppose we want to know the average weight of a black and tan coonhound.
- Somehow we're able gather data on every single one of them and find a population mean.

The sampling distribution

- It's unlikely we're going to get data on every living black and tan coonhound. Let's suppose then that we take a sample of 1,000 and find a mean weight of 55 lbs.
- Given this sample average, how are we to know how close we are (aren't) to the true population average?
- Really, we can't know. What we *can* do is to take more samples from the target population.
- If we were to record the mean from each sample collected and then graph these means in a distribution, we would get a "sampling distribution."

The central limit theorem[1]

- Theoretically, we could take an infinite number of samples of coonhounds and add their sample means to the sampling distribution.
- As we collect more samples, something special happens. First, the sampling distribution becomes a symmetric bell-curve, regardless of the shape of the underlying population distribution.
- Second, the mean of the sampling distribution begins to approximate the mean of the population.

[1] Cute video explaining the central limit theorem here: `t.ly/eBhb`.

Generalizing from a single sample

- What we really want to do is use results from a sample to draw meaningful generalizations about the target population.
- We already have the mean of the sample, but how can we think about how close we got to the population parameter of interest?
- We can use a measure of a sample's dispersion (standard deviation) to begin addressing this issue.

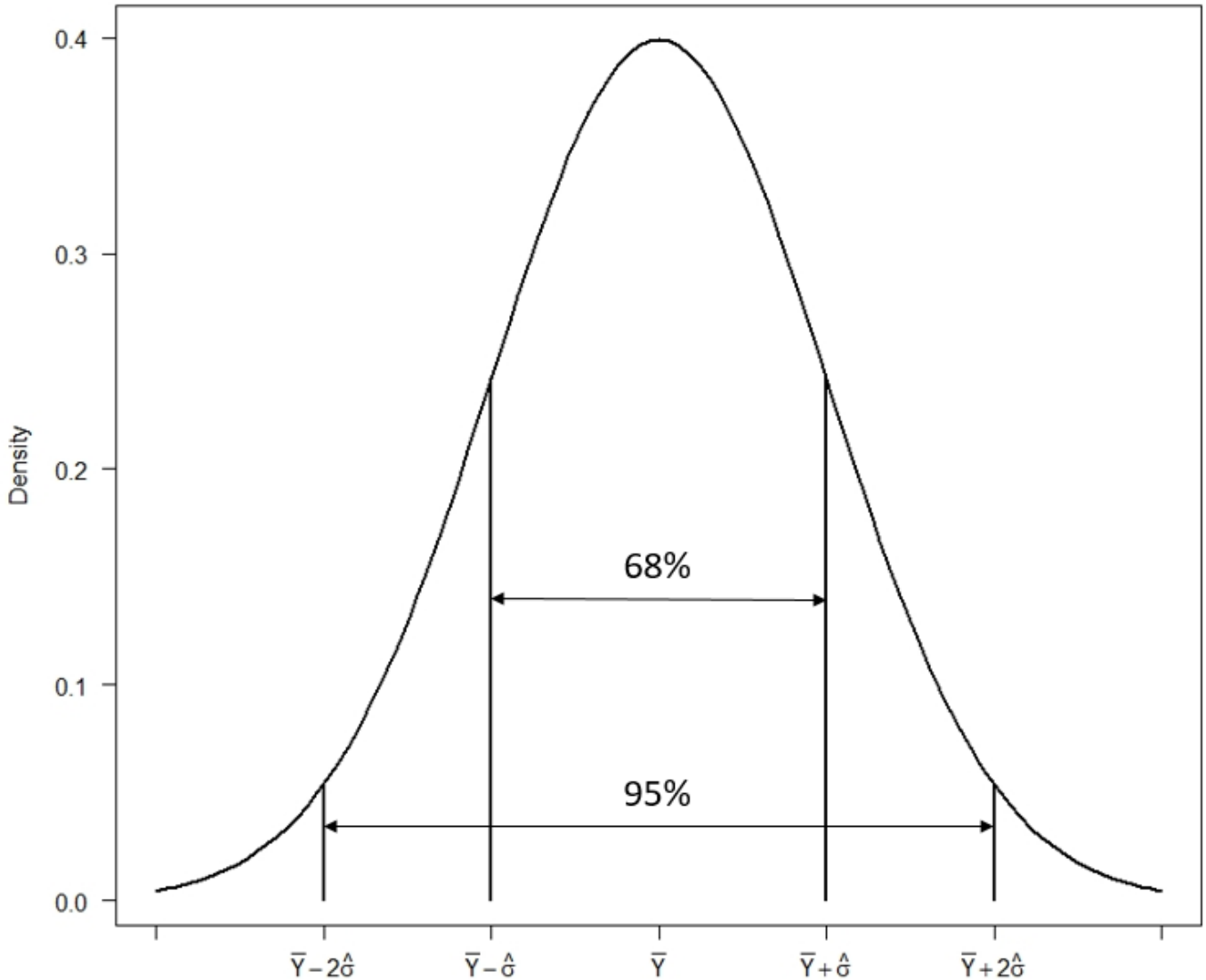
The standard error

- Remember that what we really want to do is describe a population—not so much a given sample drawn from it.
- We can use a sample's mean and dispersion to think about where a population's mean is most likely to be using a concept known as the standard error, $\hat{\sigma}$:

$$\hat{\sigma} = \frac{\sigma}{\sqrt{n}},$$

where σ is the standard deviation of the sample, and n is the sample size.

Uncertainty and the z-distribution



Characterizing uncertainty: Confidence intervals

- Because our individual samples come with random error, our estimates are mere approximations of the true population parameters of interest.
- Confidence intervals help us to describe our degree of uncertainty surrounding the true population parameter.
- Confidence intervals straddle point estimates and give an upper and lower-boundary on where the true population parameter is likely to be within a given range of confidence.

The confidence interval (cont'd.)

- Confidence intervals vary in their degrees of certainty, but the industry standard is a 95% confidence interval.
- We can interpret a 95% confidence interval as such: "Upon repeated sampling, 95% of samples' confidence intervals will contain the true population parameter of interest."
- A 95% confidence interval is computed as such:

$$C.I._{.95} \quad = \quad \bar{y} \pm \hat{\sigma}(1.96),$$

where 1.96 is plus-or-minus the number of standard errors from $\bar{y}$ in the z-distribution that summarizes 95% of all observations.

Practice: Sample mean and the confidence interval

- Okay, suppose I take a sample of 5 black and tan coonhounds, measure their weight, and observe: $\boldsymbol{y} = \{40, 45, 50, 55, 60\}$.
- Let's calculate the mean, standard deviation, standard error, and a 95% confidence interval, interpreting each as we go.

Proportions in survey research

- When we poll respondents' opinions, we are more often than not dealing with proportions.
- For example, do respondents approve of the job the president is doing (coded "1" if yes, "0" else)?
- When dealing with proportions, we often calculate the standard error as:

$$\hat{\sigma}_\rho = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

where $\hat{p}$ represents the proportion of recorded "1"s.

Proportions and the margin of error

- In polling, the margin of error is akin to the confidence interval.
- For a poll in which we want 95% confidence, we calculate the margin of error for the sample proportion as:

$$MOE_{95} = 1.96 \times \sqrt{\frac{\hat{p}^*(1-\hat{p}^*)}{n}}.$$

Proportions and the margin of error (cont'd.)

- By custom, we use $\hat{p}^* = 0.5$ regardless of our observed $\hat{p}$ to reflect greater uncertainty.
- Thus, if we had $n = 1,000$, our MOE would be 0.03.
- Suppose we polled presidential approval and found $0.\hat{4}0$ with $n = 1,000$.
- Then $MOE = 0.03$, and we'd have 95% confidence that the proportion of the public that supports the president is in the interval, 0.37 to 0.43.

Practice: Proportions and the margin of error

- Suppose we poll 10 people regarding their approval of the job the AUM chancellor is doing.
- Everyone who approves is coded "1", all who disapprove is coded "0" (ignore DKs for now) and observe: $\boldsymbol{y} = \{1, 1, 1, 1, 1, 1, 0, 0, 0, 0\}$.
- Let us calculate $\hat{p}$, $\hat{\sigma}_\rho$, and the 95% MOE.

Probability samples

- Anytime we draw samples, we'd like to know the probability that a given individual was chosen for observation.
- This is known as probability sampling and is the preferred way to make precise generalizations.

Simple random samples

- Random sampling methods are generally preferred as they reduce systematic bias.
- The most straightforward way to take a random sample is to conduct a simple random sample (SRS).
- This means that everyone in the sampling frame has an identical probability ($p = \frac{1}{n}$) of being sampled. How can we do this?

Interval samples

- Interval sampling is akin to simple random sampling. So long as there isn't structure to your chosen interval, results will approximate a SRS.
- For an interval sample, select an interval length, $1 \leq k \leq n$. Array your subjects, and sample every k^{th} observation.
- For example, you might stand outside the cafeteria from open to close and ask every 20^{th} person their opinion about the service.

Other probability samples

- Sometimes, conducting a SRS isn't feasible. This could be because the target population is difficult to reach. For example, how do we construct a sampling frame for people who have recently been hospitalized?
- Sometimes, SRS can produce samples that aren't as precise as other methods. For example, suppose I know the proportion of my target population that is male vs. female. I might want to consider a sampling strategy that will replicate that proportion in my sample.

Stratified samples

- "Strata" are discrete, mutually exclusive, exhaustive groupings in a population from which we can draw samples.
- Statified sampling, then, involves taking a series of SRSs across the chosen number of strata (here proportionate sampling may be desirable).
- It can also involve multiple layers of strata. That is, strata within strata.

Cluster samples

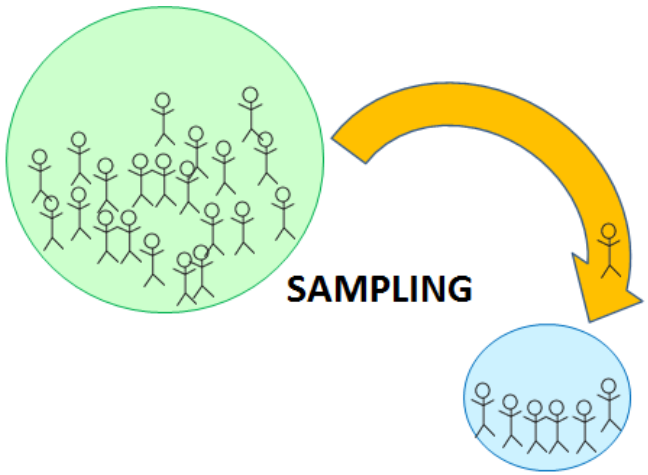
- With stratified sampling, we took SRSs from each of the strata/substrata. What if this isn't feasible, though?
- For example, it might not be possible to sample from each of the strata within a population, but it could be feasible to take a sample.
- This is cluster sampling. We identify clusters (very similar to strata); take a SRS of clusters; then study the subjects within clusters.

Non-probability sampling

- There can be compelling (and not-so-compelling) reasons why probability sampling is impractical.
- With non-probability sampling, we *do not* know the underlying probability elements are chosen for study, which can lead to bias.

Non-probability sampling methods

- Convenience sampling
- Quota sampling
- Snowball sampling



Weighting survey results

- Generally, unless we have a large sample generated by probability sampling methods, we will need to weight survey results so that they reflect the target population.
- Weighting can also be of good use if we intentionally over-sampled hard to identify sub-groups like racial minorities.

Weighting basics

- Simply put, a weight is a value given to each individual in the dataset which indicates how much they "count" in the analysis.
- Theoretically, these weights can range between 0 and infinity.
- A record assigned any weight strictly between 0 and 1 is diminished overall. Weights greater than 1 are magnified, and weights of 1 aren't affected.
- We most commonly weight for demographic factors such as race, age, education, gender, and region.

Calculating one weight

- Suppose our target population is Alabama residents, and we want to weight by voter registration status.
- We gather a sample of respondents and find that 75.0% are registered. According to data from the US Census Bureau (`t.ly/EmGx`), 68.9% of Alabamians are actually registered to vote.
- Hence, we have too many registered voters to accurately reflect Alabama residents and need to count them for less.
- To calculate the weight for registered voters, we divide the population percent by the sample percent: $68.9/75.0 = 0.92$.

Weighting with more than one variable

- Things get trickier when we want to weight with respect to more than one variable (and we do want to do that generally).
- At the end of the day, each observation needs a single weight. That is, we can't have a bunch of different weights for each variable we think is important for weighting purposes.
- For the AUM Poll, we employe a technique known as "iterative proportional weighting" to accomplish our ends.

Why weighting matters

- We don't calculate weights just for fun, but we use them to draw inferences about populations of interest.
- Later this term, when we go over summary statistics and regression analysis, we'll discuss how our description of sample characteristics can depend so much upon weighting methods.

Where to find population estimates for weights

- The US Census Bureau has just a wealth of information on the internet.
- Specifically, you should be aware of the American Community Survey and the Current Population Survey.
- Depending on what you're wanting to weight on, consider other state and federal agencies.

Some complications common to weighting

- When to "trim" a weight that's gotten out of control (we'll use 5 as a cutoff for the survey we run later on).
- Be careful with standard errors and use software that accounts for your weighting procedure when calculating these.

Conclusion

- The generalizations we draw from our samples are only as good as they are representative of the target population.
- Generally speaking, the most reliable and precise estimates come from probability samples.
- Probability samples are those in which we can express the probability any given individual will be sampled for study.
- Nevertheless, we might have compelling reasons to employ non-probability sampling, and weighting can help us to overcome some of the limitations that come with this approach.